REMARKS

In the Final Action dated February 28, 2006, claims 37-45 and 47-50 are pending and under consideration. Claims 37, 42, 43 and 47 are allowed. Claims 41 and 45 are rejected under 35 U.S.C. §112, second paragraph, as indefinite. Claims 38-41, 44 and 48-50 are rejected under 35 U.S.C. §112, first paragraph, for allegedly failing to satisfy the enablement requirement. Claims 38, 41, 44, and 48 are separately rejected under 35 U.S.C. §112, first paragraph, for allegedly failing to satisfy the written description requirement. Claims 38, 44 and 48 are rejected under 35 U.S.C. §102(b) as anticipated by Kiyoshi et al. (U.S. Patent No. 5,453,491).

This Response addresses each of the Examiner's rejections. Applicants therefore respectfully submit that the present application is in condition for allowance. Favorable consideration of all pending claims is therefore respectfully requested.

Claims 41 and 45 are rejected under 35 U.S.C. §112, second paragraph, as indefinite for reciting the terms "mature form" and "soluble form".

Initially, Applicants draw the Examiner's attention to the amendments to claims 41 and 45. Claim 41 has been amended to depend from claims 38 and 40, instead of claims 37-40; and claim 45 has been amended to depend from claims 38-40, instead of claim 37.

The Examiner contends that the metes and bounds of a "soluble form" and a "mature form" of the human NR4 molecule can only be determined by knowing at which particular amino acid residue of SEQ ID NO: 4 the designated forms start and at what residue they end. The Examiner argues that even if one skilled in the art could obtain some information regarding the domains of human NR4 based on Figure 1, it is not disclosed in the application whether the NR4 polypeptide contains more than one extracellular domain, transmembrane domain or cytoplasmic

domain. In addition, the Examiner contends that the term "mature" is ambiguous as to whether it relates to the activity of the polypeptide or the length of the polypeptide. In sum, the Examiner concludes that the information provided in the specification is insufficient for those skilled in the art to make a determination as to what constitutes a mature form, or a soluble form, of human NR4.

Applicants respectfully submit that based on the instant disclosure, including Figures 1 and 7 and the description of these figures, those skilled in the art would clearly understand that the NR4 polypeptide contains one extracellular domain, one transmembrane domain and one cytoplasmic domain.

Further, Applicants respectfully submit that NR4 is clearly identified as a haemopoietin receptor in the specification. Haemopoietin receptors include the LIF receptor, IL-6 receptor and the G-CSF receptor. The basic structure of haemopoietin receptors was known prior to the present invention and is illustrated in Plate 2 and Plate 3 (**Exhibit 1**) of the "Guidebook to Cytokines and their Receptors", Nicos A. Nicola, A Sambrook and Tooze Publication, Editors, Oxford University Press, Oxford, New York, Tokyo 1994. Exhibit 1 shows that haemopoietin receptors have a single extracellular domain, transmembrane domain and cytoplasmic domain.

As to the term "mature form", Applicants respectfully submit that the term represents a form of the NR4 polypeptide that is different in length from the newly translated NR4 polypeptide as a result of post-translational processing. In the context of the present invention, the "mature form" of the NR4 protein can be glycosylated or unglycosylated, depending on the expression system employed in the recombinant production of the protein.

Applicants disagree with the Examiner's conclusion that a "soluble form" and a "mature form" of human NR4 can not be sufficiently defined without a specific disclosure in the specification of the precise starting and ending amino acid residues of the respective forms. Applicants respectfully submit that based on the information provided in the specification, those skilled in the art would be able to make a determination of the starting and ending amino acid residues, or at least the approximate starting and ending positions, and to readily confirm the accuracy of such a determination by routine experimentation.

Applicants submit that the detailed deconstruction of the murine NR4 (IL-13 receptor alpha chain) in Example 6 clearly demonstrates that appropriate means were available in 1995/1996 to one skilled in the art to determine the signal sequence and trans-membrane regions of a protein. Through these means, coupled with the disclosures relating to the analysis and comparison of the murine and human NR4 sequences, those skilled in the art would have been able to determine the signal peptide and transmembrane region of human NR4, at the time the present application was filed.

In this connection, Applicants submit that there are a number of publications that were available to one skilled in the art describing how to identify the signal peptide and transmembrane regions of proteins, as well as a number of programs available on the web.

For example, Gunnar von Heijne described a method for predicting the site of cleavage between a signal sequence and the mature protein in 1986 using a weight-matrix approach. See, "*A new method for predicting signal sequence cleavage sites*" Gunnar von Heijne, *Nucleic Acid Research*, 14(11): 4683-4690, 1986 (**Exhibit 2**). This was the most widely

used method for predicting the location of the cleavage site of signal peptides in 1995/1996. This was in fact the common method at the time for predicting signal peptide cleavage points.

A paper published in 1999 reviews the PROSITE database developed in 1998 that consists of biologically significant patterns and profiles formulated in such a way that, with the appropriate computational tools, it can help determine to which known family of proteins a new sequence belongs to or which known domains it contains. The database was developed to enable the identification and function of uncharacterized proteins translated from genomic or cDNA sequences. See, "The PROSITE database, its status in 1999", Hoffmann et al., *Nucleic Acid Research*, 27(1): 215-219, 1999 (**Exhibit 3**).

Further, Applicants submit that the transmembrane region was commonly determined by hydrophobicity analysis. For example, the following two papers described strategies for predicting transmembrane topology of prokaryotic and eukaryotic membrane proteins.

"Membrane Protein Structure Prediction, Hydrophobicity analysis and the positive inside rule", Gunnar von Heijne, *Journal of Moelcular Biology*, 225: 487-494, 1992 (abstract attached as **Exhibit 4**); and "Predicting the Topology of eukaryotic membrane proteins"; Sipos L. and von Heijne G., *Eur J. Biochem*, 213(3): 1333-1340, 1993 (abstract attached as **Exhibit 5**).

In addition, there is also a website for "Tmpred", a program that makes a prediction of membrane-spanning regions and their orientation. See www.ch.embnet.org/software/TMPRED form.html. The algorithm is based on the statistical analysis of TMbase, a database of naturally occurring transmembrane proteins. The prediction is made using a combination of several weight-matrices for scoring. TMbase was published in 1993 (see, K. Hofmann & W. Stoffel,

"TMbase - A database of membrane spanning protein segments", *Biol. Chem.*, Hoppe-Seyler, 374: 166, 1993, **Exhibit 6**).

Accordingly, Applicants respectfully submit that a "soluble form" and a "mature form" of human NR4 are sufficiently defined in the specification, despite the absence of a specific disclosure of the precise starting and ending amino acid residues of the respective forms. In view of the foregoing, Applicants respectfully submit that the rejection of claims 41 and 45 under 35 U.S.C. §112, second paragraph, is overcome. Withdrawal of the rejection is respectfully requested.

Claims 38-41, 44 and 48-50 are rejected under 35 U.S.C. §112, first paragraph, for allegedly failing to satisfy the enablement requirement. Claims 38, 41, 44, and 48 are separately rejected under 35 U.S.C. §112, first paragraph, for allegedly failing to satisfy the written description requirement. The Examiner's principal concern appears to be directed to the recitation of "a part or fragment of SEQ ID NO: 4" in independent claim 38. Essentially, the Examiner contends that the recited "part or fragment of SEQ ID NO: 4" is not defined by any structural or functional feature.

Applicants respectfully submit that the specification does provide guidance for "parts" and "fragments" of NR4. For example, Example 6 of the specification (page 37 and Figure 1) defines the various domains of murine NR4 including a signal sequence, an extracellular domain, a transmembrane domain, and a cytoplasmic domain. Example 11 (pages 39-40) discloses that SEQ ID NO: 4 is the human homolog of murine NR4 with 75% similarity at the amino acid level; and Figure 7 aligns the human sequence with the murine sequence. In view of the disclosure in the specification, the term "part or fragment" of SEQ ID NO: 4, as presently recited,

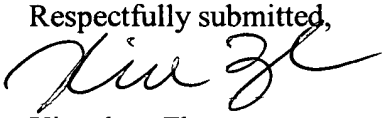
is fully supported by the specification. As such, the enablement and written description rejections under 35 U.S.C. §112, first paragraph, are overcome. Withdrawal of the rejections is respectfully requested.

Claims 38, 44 and 48 are rejected under 35 U.S.C. §102(b) as anticipated by Kiyoshi et al. (U.S. Patent No. 5,453,491). The rejection is apparently made based on the Examiner's interpretation of the term "a part" of SEQ ID NO: 4 as reading on an amino acid, which is disclosed by Kiyoshi et al.

Applicants respectfully submit that in light of the specification, those skilled in the art would not interpret the term "part" of SEQ ID NO: 4 to include simply an amino acid. As described in the specification, a derivative of an NR4 molecule, which includes parts or fragments of an NR4 molecule, can be a functional molecule, e.g., capable of binding to IL-13, or a non-functional but immunogenic molecule. See page 7, lines 1-13 of the specification. Those skilled in the art would not consider a single amino acid as "a part" of SEQ ID NO: 4. Accordingly, the §102(b) rejection based on Kiyoshi et al. is overcome and withdrawal thereof is respectfully requested.

In view of the foregoing amendments and remarks, it is firmly believed that the subject application is in condition for allowance, which action is earnestly solicited.

Respectfully submitted,



Xiaochun Zhu
Registration No. 56,311

Scully, Scott, Murphy & Presser, P. C.
400 Garden City Plaza-STE 300
Garden City, NY 11530
(516) 742-4343
XZ:ab
Encls.: Exhibits 1-6

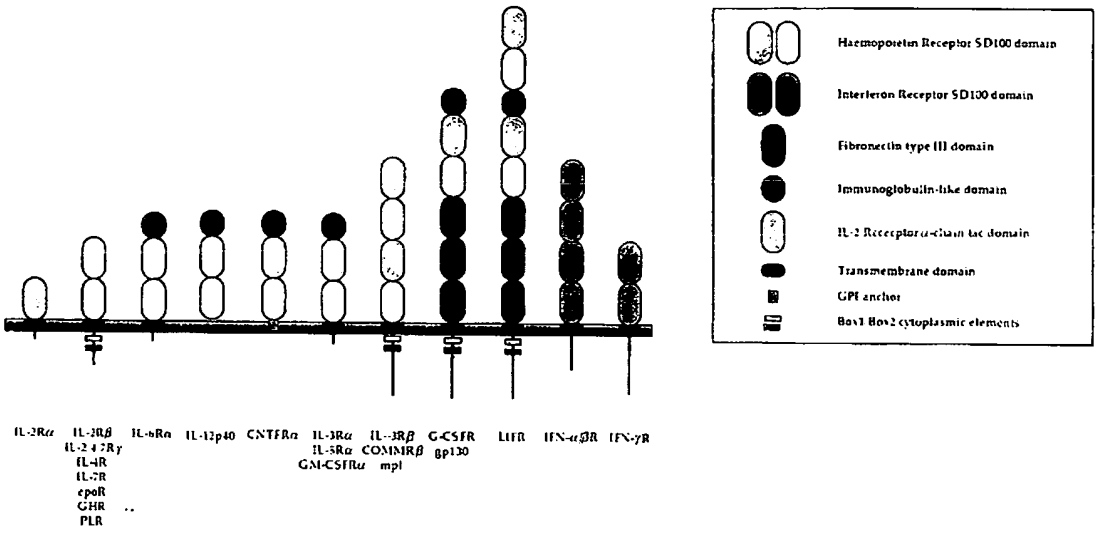


Plate 2. Schematic view of the haemopoietin/interferon receptor family.
Abbreviations: IL = interleukin, epo = erythropoietin, GH = growth hormone, PL = prolactin, CNTF = ciliary neurotrophic factor, G-CSF = granulocyte colony-stimulating factor, GM-CSF = granulocyte-macrophage colony-stimulating factor, COMM = common receptor chain to IL-3, IL-5, and GM-CSF, LIF = leukaemia inhibitory factor, IFN = interferon, R = receptor.
(This figure was adapted from Figure 1 of Gearing and Ziegler 1993.)

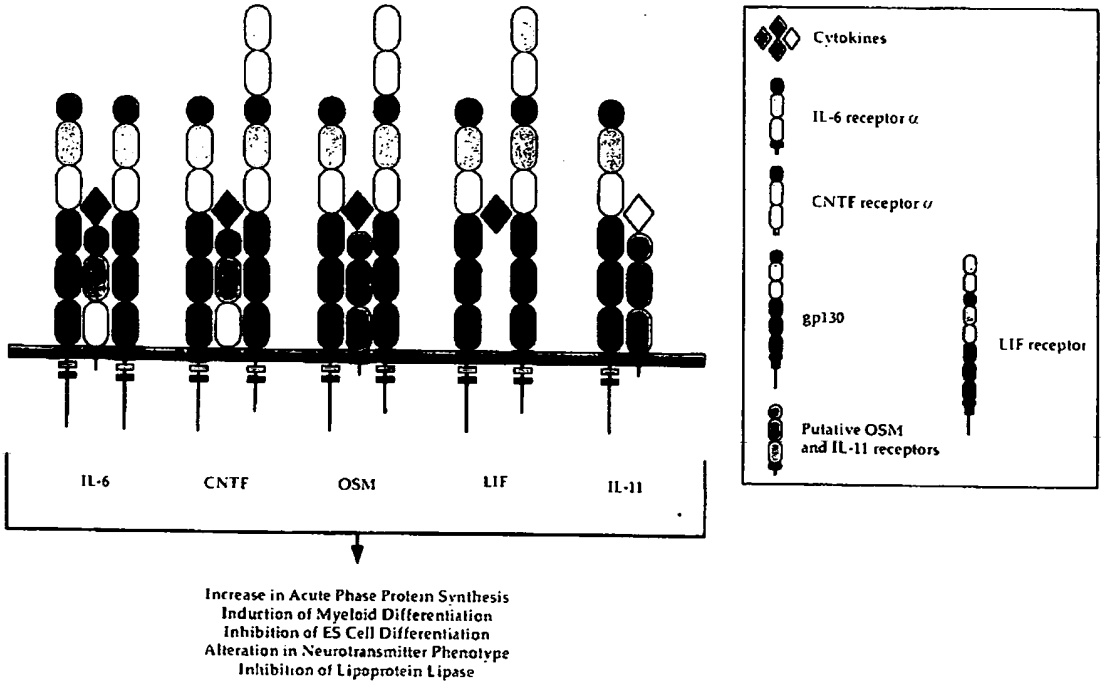
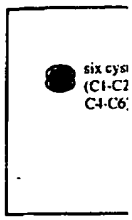


Plate 3. Shared receptor subunits may contribute to cytokine pleiotropy.
Abbreviations: IL = interleukin, OSM = oncostatin M, CNTF = ciliary neurotrophic factor, LIF = leukaemia inhibitory factor.



NGFR p7

Plate 4. Scl
Abbreviation
homologue,
R = receptor
(This figure

Volume 14 Number 11 1986 Nucleic Acids Research

A new method for predicting signal sequence cleavage sites

Gunnar von Heijne

Research Group for Theoretical Biophysics, Department of Theoretical Physics, Royal Institute of Technology, S-100 44 Stockholm, Sweden

Received 5 March 1986; Revised and Accepted 5 May 1986

ABSTRACT

A new method for identifying secretory signal sequences and for predicting the site of cleavage between a signal sequence and the mature exported protein is described. The predictive accuracy is estimated to be around 75-80% for both prokaryotic and eukaryotic proteins.

INTRODUCTION

The transient N-terminal signal sequence found on most secretory proteins serves to initiate export across the inner membrane (in prokaryotes) or the endoplasmic reticulum (in eukaryotes). Three structurally and, possibly, functionally distinct regions have been identified as the basic building-blocks of a secretory signal sequence: a basic N-terminal region (n-region), a central hydrophobic region (h-region), and a more polar C-terminal region (c-region) (1). The structural determinants for cleavage of the signal sequence from the mature protein once export is under way seems to reside in the n- and h-regions, with positions -3 and -1 relative to the cleavage site being the most important ones (2,3). Indeed, this "(-3,-1)-rule" has been used quite successfully to predict the most likely site of cleavage directly from the primary sequence (2).

In view of the great interest in secretory proteins and the fact that most such proteins are known only from their DNA sequence, it is important to assess and, if possible, to improve upon the predictive accuracy of the original method. In this paper, I present a new scheme based on a weight-matrix approach that can be expected to give correct predictions about 75-80% of the time when applied to new sequences (both prokaryotic and eukaryotic). This represents a substantial gain over the old method, which is shown to be around 65% and 45% accurate for eukaryotic and prokaryotic proteins, respectively.

METHODS

161 eukaryotic and 36 prokaryotic non-homologous signal sequences with known cleavage sites were chosen from my collection of signal sequences totalling at the present time some 450 eukaryotic and 80 prokaryotic entries. The prokaryotic sample did not include any sequences known to be cleaved by the lipoprotein signal peptidase (signal peptidase II) (4).

Weight-matrices W(a,i) (see below) were calculated from the observed amino acid counts in each position, N(a,i), (i.e. the number of residues of type a in position i) with all sequences aligned from their known site of cleavage between positions -1 and +1, by first dividing all counts by their respective expected abundance in proteins in general, <N(a)> (Tables 1 & 2, last column), and then taking the natural logarithms of these quotients: $W(a,i) = \ln(N(a,i)/\langle N(a)\rangle)$. To correct for the limited size of the data base, all zero-elements in the amino acid count matrices were put equal to one before the divison. Zero-counts in positions -3 and -1 were treated differently: they were also put equal to one, but then divided by the total number of sequences in the sample, N, rather than the expected number of residues, e.g. $W(a,-1) = \ln(1/N)$ if $N(a,-1) = 0$.

The most probable cleavage site was identified by scanning the sequence in question with the appropriate weight-matrix and summing the weights for each position, i.e. $S(i) = W(a_{i-p},i-p) + W(a_{i-p+1},i-p+1) + \ldots + W(a_{i+q},i+q)$ where the summation window extends from position i-p to i+q. The predicted cleavage site j is the one with the highest S-value, $S(j) = \max[S(i);\ i=1-p,\ldots,L-q]$, where L is the length of the sequence analyzed. As shown below, maximum predictive accuracy was obtained for p=-12 and q=2.

RESULTS

The (-3,-1)-rule

Based on previous statistics (2), acceptable cleavage sites were suggested to conform to the following rules: the residue in position -1 must be small, i.e. either Ala, Ser, Gly, Cys, Thr, or Gln; the residue in position -3 must not be aromatic (Phe, His, Tyr, Trp), charged (Asp, Glu, Lys, Arg), or large and polar (Asn, Gln). Further, it was suggested that Pro must be absent from positions -3 through +1. The new amino acid counts presented in Tables 1 & 2 are based on more than twice as many sequences; nevertheless, the (-3,-1)-rule is seen to hold remarkably well. The only exceptions found to date among eukaryotic proteins are one sequence with Leu in -1, one with Pro in -2, and three with Pro in -1. Thus, barring sequencing errors, we must

Table 1 **Amino acid counts for eukaryotic signal sequences**
The average composition (last column) is from Ref.(10)

	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	Expected
A	16	13	14	15	20	18	18	17	25	15	47	6	80	18	6	14.5
C	3	6	9	7	9	14	6	8	5	6	19	3	9	8	3	4.5
D	0	0	0	0	0	0	0	0	5	3	0	5	0	10	11	8.9
E	0	0	0	1	0	0	0	0	3	7	0	7	0	13	14	10.0
F	13	9	11	11	6	7	18	13	4	5	0	13	0	6	4	5.6
G	4	4	3	6	3	13	3	2	19	34	5	7	39	10	7	12.1
H	0	0	0	0	0	1	1	0	5	0	0	6	0	4	2	3.4
I	15	15	8	6	11	5	4	8	5	1	10	5	0	8	7	7.4
K	0	0	0	1	0	0	1	0	0	4	0	2	0	11	9	11.3
L	71	68	72	79	78	45	64	49	10	23	8	20	1	8	4	12.1
M	0	3	7	4	1	6	2	2	0	0	0	1	0	1	2	2.7
N	0	1	0	1	1	0	0	0	3	3	0	10	0	4	7	7.1
P	2	0	2	0	0	4	1	8	20	14	0	1	3	0	22	7.4
Q	0	0	0	1	0	6	1	0	10	8	0	18	3	19	10	6.3
R	2	0	0	0	0	1	0	0	7	4	0	15	0	12	9	7.6
S	9	3	8	6	13	10	15	16	26	11	23	17	20	15	10	11.4
T	2	10	5	4	5	13	7	7	12	6	17	8	6	3	10	9.7
V	20	25	15	18	13	15	11	27	0	12	32	3	0	8	17	11.1
W	4	3	3	1	1	2	6	3	1	3	0	9	0	2	0	1.8
Y	0	1	4	0	0	1	3	1	1	2	0	5	0	1	7	5.6

admit the possibility that residues other than the classical (-3,-1)-kinds can be used in position -1, but only when no better cleavage site is available in the vicinity (this is true for all five exceptions).

A few other points can also be made. First, the constraints on the prokaryotic sequences in the (-3,-1)-region seem even stronger than for the eukaryotic ones: only Ala, Gly, Ser and Thr have been found in -1, and only Ala, Gly, Leu, Ser, Thr, and Val in -3. Second, Leu is abundant in the prokaryotic sample up to and including position -8, but its incidence drops precipitously in position -7, where it is replaced by the likewise hydrophobic but less strongly helix-inducing residues Val and Phe. Only from position -6 do we find predominantely polar residues. Finally, there is a notable imbalance between the basic residues Arg and Lys in the c-region of the eukaryotic signal sequences, with 26 Arg and only 6 Lys (Arg/Lys = 4.3). This is in sharp contrast to the n-region where Arg/Lys = 66/72 = 0.9 and to proteins in general where the expected ratio is 0.6 (Table 1, last column).

Table 2 **Amino acid counts for prokaryotic signal sequences**
The average composition (last column) is from Ref.(10)

	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2	Expected
A	10	8	8	9	6	7	5	6	7	7	24	2	31	18	4	3.2
C	1	0	0	1	1	0	0	1	1	0	0	0	0	0	0	1.0
D	0	0	0	0	0	0	0	0	0	0	0	0	0	2	8	2.0
E	0	0	0	0	0	0	0	0	0	0	0	1	0	4	8	2.2
F	2	4	3	4	1	1	8	0	4	1	0	7	0	1	0	1.3
G	4	2	2	2	3	5	2	4	2	2	0	2	2	1	0	2.7
H	0	0	1	0	0	0	0	1	1	0	0	7	0	1	0	0.8
I	3	1	5	1	5	0	1	3	0	0	0	0	0	0	2	1.7
K	0	0	0	0	0	0	0	0	0	1	0	2	0	3	0	2.5
L	8	11	9	8	9	13	1	0	2	2	1	2	0	0	1	2.7
M	0	2	1	1	3	2	3	0	1	2	0	4	0	0	1	0.6
N	0	0	0	0	0	0	0	1	1	1	0	3	0	1	4	1.6
P	0	1	1	1	1	1	2	3	5	2	0	0	0	0	5	1.7
Q	0	0	0	0	0	0	0	0	2	2	0	3	0	0	1	1.4
R	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1.7
S	1	0	1	4	4	1	5	15	5	8	5	2	2	0	0	2.6
T	2	0	4	2	2	2	2	2	5	1	3	0	1	1	2	2.2
V	5	7	1	3	1	4	7	0	0	4	3	0	0	2	0	2.5
W	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0.4
Y	0	0	0	0	0	0	0	0	0	3	0	1	0	0	0	1.3

Construction of weight-matrices

Weight-matrix methods have been used for a number of years to locate signals in nucleic acid sequences (see (5) for a thorough discussion). Their use for pattern recognition in protein sequences requires a larger data base (20 amino acids rather than 4 bases must be scored for in each position), but is no different in principle. Basically, one converts the observed number of each kind of residue in each position in a sample of aligned "signals" into a measure of the probability of finding that particular kind of residue in that particular position - the probability weight-matrix - by a suitable normalization. Then, any new sequence can be scanned by a moving window (looking up the respective probabilities in the weight-matrix and multiplying together for each position of the window) to get a measure of the fit to the sample used in the construction of the weight-matrix. The highest-scoring window-position is then taken as the prediction for the location of the signal, if the score is above some minimum value.

To score for possible signal sequence function, and to locate the most probable cleavage site in a putative signal sequence, weight-matrices for prokaryotic and eukaryotic signal sequences were constructed as follows. The raw amino acid counts for the two samples (Tables 1 & 2) were divided by the expected number $\langle N(a) \rangle$ of each kind of residue given amino acid frequencies as in soluble proteins in general (last columns). Except for positions -3 and -1 relative to the cleavage site, all matrix elements with zero counts were normalized as $1/\langle N(a) \rangle$. For positions -3 and -1, where there is good reason both from previous statistical and experimental studies to believe that only a subset of all residues are allowed (2,6), the more stringent normalization $1/N$ was used for the zero-count elements (where N is the total number of sequences in the sample). The final weight-matrix was obtained by taking the natural logarithms of the normalized values, thus reducing the ensuing probability calculations to summations rather than multiplications of the weight-matrix elements.

Assessment of the predictive accuracy

When the two weight-matrices were used to predict the cleavage sites in the samples used in their construction, virtually all sites were correctly identified (87% in the eukaryotic sample, 100% in the prokaryotic sample). However, these sequences are at an advantage relative to sequences not included in the matrix: when correctly aligned with the weight-matrix, all residues in a sequence included in the weight-matrix sample will correspond to a count, and a spuriously high predictive accuracy will be found.

To avoid this problem, the eukaryotic sample was divided into 7 subsamples, each of 23 sequences. For each subsample, the remaining 138 sequences were used to construct a new weight-matrix, and this matrix was then applied to the subsample. Similarly, the prokaryotic sample was divided into 4 subsamples, each of 9 sequences. All subsequent calculations were carried out by summing the results for the subsamples.

Weight-matrices including positions -15 to +5 were first used to determine the effect of ignoring residues at either end in the predictions. It was found that positions -13 to +2 were sufficient to obtain maximal predictive accuracy (for the prokaryotic sample, positions -5 to +2 were sufficient but the full -13 to +2 range was used nevertheless): with this choice, 125 out of 161 eukaryotic and 32 out of 36 prokaryotic cleavage sites (78% and 89%) were correctly identified with a standard deviation of about ±10% in each case. For an additional 19 eukaryotic and 2 prokaryotic sequences, the correct site had the second-highest score. These values should

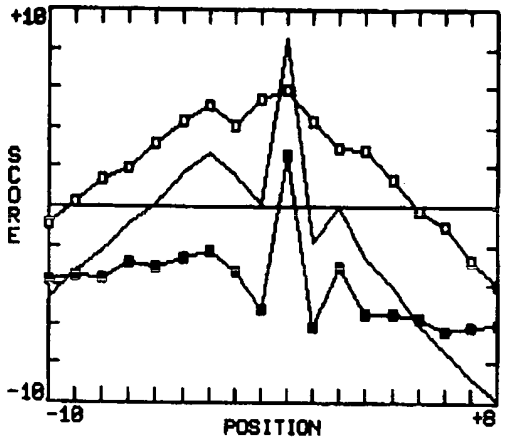


Figure 1 Average h- and c-region scores as a function of the position of the moving window. Open squares: h-region; solid squares: c-region; full line: total score.

be compared with the predictive accuracy of the older method (as implemented in a program kindly communicated by Dr. H.S. Ip, Rockefeller University). When this method was applied to the 121 sequences in the eukaryotic sample that were not included in the original statistics (2), 77/121 (64%) of the known cleavage sites were correctly identified, and only 17/36 (47%) of the prokaryotic ones were found.

With -13 to +2 weight-matrices, the contribution to the overall success from individual positions was also investigated. Only positions -3 and -1 had any strong impact; when one or the other was left out in the calculations the percentage of correctly identified eukaryotic sites dropped to 61% and 53%, respectively (81% and 69% for the prokaryotic sample).

As has been shown previously (1,7), residues -13 to -6 correspond to the h-region in the "average" eukaryotic signal sequence, residues -5 to -1 correspond to the c-region, and residues +1 and +2 seem to be selected such that few alternative cleavage sites should exist in the vicinity of the correct one (i.e. residues -5 to +2 can be included in an extended c-region). Thus, it is possible to calculate the scores for the h- and c-regions separately by summing the contributions from positions -13 to -6 and -5 to +2, respectively. As shown in Fig.1, the average h-region score for the eukaryotic sample increases slowly as the window is moved up to position -1 (the known cleavage site), and then decreases. The average c-region score

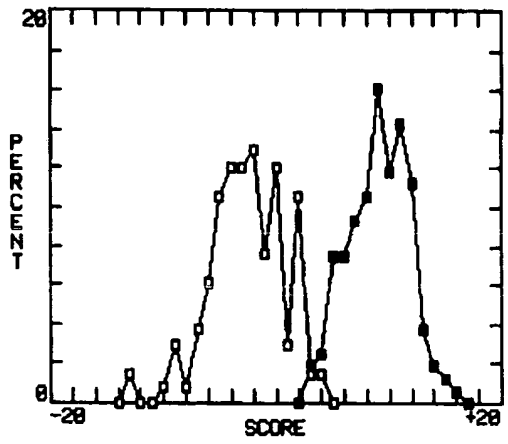


Figure 2 **Distribution of maximum scores for signal sequences and cytosolic proteins.** Open squares: cytosolic proteins; solid squares: signal sequences.

shows a more dramatic behaviour, with a pronounced peak in position -1 and troughs in positions -2 and +1, reflecting the match to the (-3,-1)-pattern and the tendency to have residues in position -2 that do not fit this pattern (see Tables 1 & 2). Similar curves are obtained for the prokaryotic sample (not shown).

Interestingly, 35 out of the 36 erroneous predictions for the eukaryotic sequences fall on the N-terminal side of the correct cleavage site, mostly in the region -6 to -3 (30/36). About half of these result from matches with a higher score in the h-region but a lower one in the c-region than calculated for the correct site, whereas only 6 out of 36 have higher c- and lower h-region scores than the correct site. I have thus tried to improve the predictive accuracy in various ways, e.g. by multiplying the -3 and -1 weigths or the whole c-region score by an extra factor, or by allowing a small variation in the distance between the h- and c-regions, but have not been able to obtain more than marginal improvements on the order of 2-4% in the overall success-rate.

The method described here not only allows prediction of the most likely cleavage site in new signal sequences, it also makes it possible to discriminate quite efficiently between putative signal sequences and the N-terminal regions of cytosolic proteins. The distribution of maximum scores for the eukaryotic signal sequences is shown in Fig.2, together with the

corresponding distribution obtained for a sample of 132 40-residues long N-teminal regions of cytosolic eukaryotic proteins (8). Only 3/161 (2%) of the signal sequences have maximum scores < 3.5; conversely, only 2/132 (2%) of the cytosolic sequences have maximum scores > 3.5. This level of discrimination compares favourably with that obtained with a recently published signal-sequence detecting algorithm (9).

DISCUSSION

Using a standard weight-matrix approach easily implemented even on a micro-computer, it is possible to set up a prediction method that (i) provides a clean discrimination between signal sequences and the N-terminal region in cytosolic proteins, and (ii) can be expected to identify the correct cleavage site 75-80% of the time when applied to new sequences not included in the data base (both prokaryotic and eukaryotic). This represents a significant improvement over previous methods.

Since the first submission of this work, another 36 eukaryotic signal sequences with known cleavage sites have been added to the data base. Using the same weight-matrix as above (Table 1), 75% of these sites were correctly predicted.

ACKNOWLEDGEMENT

This work was supported by a grant from the Swedish Natural Sciences Research Council.

REFERENCES

(1) von Heijne,G. (1985) J.Mol.Biol. 184, 99-105.
(2) von Heijne,G. (1983) Eur.J.Biochem. 133, 17-21.
(3) Perlman,D., and Halvorson,H.O. (1983) J.Mol.Biol. 167, 391-409.
(4) Mollay,C. (1985) in The Enzymology of Post-translational Modification of Proteins, Vol.2, pp. 1-23, Academic Press, London.
(5) Staden,R. (1984) Nuc.Acids Res. 12, 505-519.
(6) Kuhn,A., & Wickner,W. (1985) J.Biol.Chem. 260, 15914-15918.
(7) von Heijne,G. (1984) J.Mol.Biol. 173, 243-251.
(8) Flinta,C., Persson,B., Jörnvall,H., and von Heijne,G. (1986) Eur.J.Biochem. 154, 193-196.
(9) McGeoch,D.J. (1985) Virus Res. 3, 271-286.
(10) Klapper,H.M. (1977) Biochem.Biophys.Res.Commun. 78, 1018-1024.

Nucleic Acids Research, 1999, Vol. 27, No. 1 215–219

The PROSITE database, its status in 1999

Kay Hofmann, Philipp Bucher[1,*], Laurent Falquet[1] and Amos Bairoch[2]

MEMOREC, Stoffel GmbH, Stoeckheimer Weg 1, D-50829 Koeln, Germany, [1]Swiss Institute of Bioinformatics (SIB), Swiss Institute for Experimental Cancer Research (ISREC), CH-1066 Epalinges/Lausanne, Switzerland and [2]Swiss Institute of Bioinformatics (SIB), Department of Medical Biochemistry, University of Geneva, 1 rue Michel Servet, CH-1211 Geneva 4, Switzerland

Received October 16, 1998; Accepted October 21, 1998

ABSTRACT

The PROSITE database (http://www.expasy.ch/sprot/prosite.html) consists of biologically significant patterns and profiles formulated in such a way that with appropriate computational tools it can help to determine to which known family of protein (if any) a new sequence belongs, or which known domain(s) it contains.

BACKGROUND

PROSITE ([1,2]) is a method of identifying what is the function of uncharacterized proteins translated from genomic or cDNA sequences. It consists of a database of biologically significant patterns and profiles formulated in such a way that with appropriate computational tools it can rapidly and reliably determine to which known family of protein (if any) the new sequence belongs, or which known domain(s) it contains.

In some cases the sequence of an unknown protein is too distantly related to any protein of known structure to detect its resemblance by overall sequence alignment. However, relationships can be revealed by the occurrence in its sequence of a particular cluster of residue types, which is variously known as a pattern, motif, signature or fingerprint. These motifs arise because specific region(s) of a protein which may be important, for example, for their binding properties or for their enzymatic activity are conserved in both structure and sequence. These structural requirements impose very tight constraints on the evolution of this small but important portion(s) of a protein sequence. The use of protein sequence patterns or profiles to determine the function of proteins is becoming very rapidly one of the essential tools of sequence analysis. Many authors ([3,4]) have recognized this reality. Based on these observations, we decided in 1988, to actively pursue the development of a database of regular expression-like patterns, which would be used to search against sequences of unknown function.

But, while sequence patterns are very useful, there are a number of protein families as well as functional or structural domains that cannot be detected using patterns due to their extreme sequence divergence. Typical examples of important functional domains, which are weakly conserved, are the globins, the immunoglobulin, and the SH2 and SH3 domains. In such domains there are only a few sequence positions which are well conserved. Any attempt to build a consensus pattern for such regions will either fail to pick up a significant proportion of the protein sequences that contain such a region (false negatives) or will pick up too many proteins that do not contain the region (false positives).

The use of techniques based on profiles or weight matrices (the two terms are used synonymously here) allows the detection of such proteins or domains. A profile is a table of position-specific amino acid weights and gap costs. These numbers (also referred to as scores) are used to calculate a similarity score for any alignment between a profile and a sequence, or parts of a profile and a sequence. An alignment with a similarity score higher than or equal to a given cut-off value constitutes a motif occurrence. As with patterns, there may be several matches to a profile in one sequence, but multiple occurrences in the same sequences must be disjoint (non-overlapping) according to a specific definition included in the profile. Another feature that distinguishes patterns from profiles is that the latter are usually not confined to small regions with high sequence similarity. Rather they attempt to characterize a protein family or domain over its entire length.

We therefore started in 1994 to complement the approach based on patterns by gradually adding to PROSITE profile entries. The profile structure ([5,6]) used in PROSITE is similar to but slightly more general than the one introduced by Gribskov and co-workers ([7]); additional parameters allow representation of other motif descriptors, including the currently popular hidden Markov models ([8]). Profiles can be constructed by a large variety of different techniques. The classical method developed by Gribskov and co-workers ([9]) requires a multiple sequence alignment as input and uses a symbol comparison table to convert residue frequency distributions into weights. Most profiles included in PROSITE are generated by this procedure applying recently described modifications ([10,11]). In some cases we also applied alternative profile construction methods including structure-based approaches and methods involving hidden Markov modelling.

LEADING CONCEPTS

The design of PROSITE follows five leading concepts.

Completeness. For such a compilation to be helpful in the determination of protein function, it is important that it contains as many biologically meaningful patterns and profiles as possible.

*To whom correspondence should be addressed. Tel: +41 21 692 5892; Fax: +41 21 652 6933; Email: philipp.bucher@isrec.unil.ch

1a) A documentation (textbook) entry from the PROSITE.DOC file

```
{PDOC00043}
{PS00041; HTH_ARAC_FAMILY_1}
{PS01124; HTH_ARAC_FAMILY_2}
{BEGIN}
*************************************************************
* Bacterial regulatory proteins, araC family signature and profile *
*************************************************************

The many bacterial transcription regulation proteins which bind DNA through a
'helix turn-helix' motif can be classified into subfamilies on the basis of
sequence similarities. One of these subfamilies groups together the following
proteins [1,2]:

- aarP, a transcriptional activator of the 2' N acetyltransferase gene in
  Providencia stuartii.
- ada, an Escherichia coli and Salmonella typhimurium bifunctional protein
  that repairs alkylated guanine in DNA by transferring the alkyl group at
  the O(6) position to a cysteine residue in the enzyme. The methylated
  protein acts a positive regulator of its own synthesis and of the alkA,
  alkB and aidB genes.
  adeA, a Bacillus subtilis bifunctional protein that acts both as a
  transcriptional activator of the ada operon and as a methylphosphotriester-
  DNA alkyltransferase.
  adiT, an Escherichia coli protein of unknown function.
- aggR, the transcriptional activator of aggregative adherence fimbria I
  expression in enteroaggregative Escherichia coli.
- appY, a protein which acts as a transcriptional activator of acid
  phosphatase and other proteins during the deceleration phase of growth and
  acts as a repressor for other proteins that are synthesized in exponential
  growth or in the stationary phase.
  araC, the arabinose operon regulatory protein, which activates the
  transcription of the araBAD genes.
- cafR, the Yersinia pestis F1 operon positive regulatory protein.
- celD, the Escherichia coli cel operon repressor.
  cfaD, a protein which is required for the expression of the CFA/I adhesin
  of enterotoxigenic Escherichia coli.
  cmvR, a transcriptional activator of fimbrial genes in enterotoxigenic
  Escherichia coli.
  envY, the porin thermoregulatory protein, which is involved in the control
  of the temperature-dependent expression of several Escherichia coli
  envelope proteins such as ompF, ompC, and lamB.
- exsA, an activator of exoenzyme S synthesis in Pseudomonas aeruginosa.
- fapR, the positive activator for the expression of the 987P operon coding
  for the fimbrial protein in enterotoxigenic Escherichia coli.
  hrpB, a positive regulator of pathogenicity genes in Burkholderia
  solanacearum.
  invF, the Salmonella typhimurium invasion operon regulator.
- marA, which may be a transcriptional activator of genes involved in the
  multiple antibiotic resistance (mar) phenotype.
- melR, the melibiose operon regulatory protein, which activates the
  transcription of the melAB genes.
- mixE, a Shigella flexneri protein necessary for secretion of ipa invasins.
  mmsR, the transcriptional activator for the mmsAB operon in Pseudomonas
  aeruginosa.
- msmR, the multiple sugar metabolism operon transcriptional activator in
  Streptococcus mutans.
- pchR, a Pseudomonas aeruginosa activator for pyochelin and ferripyochelin
  receptor.
  perA, a transcriptional activator of the eaeA gene for intimin in
  enteropathogenic Escherichia coli.
- pocR, a Salmonella typhimurium regulator of the cobalamin biosynthesis
  operon.
  pqrA, from Proteus vulgaris.
  raiR, the regulator of the raffinose operon in Pediococcus pentosaceus.
- ramA, from Klebsiella pneumoniae.
  rhaR, the Escherichia coli and Salmonella typhimurium L rhamnose operon
  transcriptional activator.
- rhaS, an Escherichia coli and Salmonella typhimurium positive activator of
  genes required for rhamnose utilization.
  rmA, a protein which is required for the expression of the col and csz
  adhesins of enterotoxigenic Escherichia coli
- rob, a protein which binds to the right arm of the replication origin oriC
  of the Escherichia coli chromosome.
- soxS, a protein that, with the soxR protein, controls a superoxide response
  regulon in Escherichia coli.
- tetD, a protein from transposon Tn10
- tcpH or toxT, the Vibrio cholerae transcriptional activator of the tcp
  operon involved in pilus biosynthesis and transport.
  thcR, a probable regulator of the thc operon for the degradation of the
  thiocarbamate herbicide EPTC in Rhodococcus sp. strain N186/21.
  ureR, the transcriptional activator of the plasmid encoded urease operon in
  Enterobacteriaceae.
  virF and lcrF, the Yersinia virulence regulon transcriptional activator.
- virF, the Shigella transcriptional factor of invasion related antigens
  spaBCD.
- xylR, the Escherichia coli xylose operon regulator.
  xylS, the transcriptional activator of the Pseudomonas putida TOL plasmid
  (pWW0. pWW3) and pDK1) meta operon (xylDLEGF genes).
- yieG, an Escherichia coli hypothetical protein.
  yhiW, an Escherichia coli hypothetical protein
  yhiX, an Escherichia coli hypothetical protein.
- ydiL, an Escherichia coli hypothetical protein.
- yijO, an Escherichia coli hypothetical protein.
- yuxC, a Bacillus subtilis hypothetical protein.
- yxlC, a Bacillus subtilis hypothetical protein.

Except for celD, all of these proteins seem to be positive transcriptional
factors. Their size range (from 107 (soxS) to 529 (ydbC) residues.

The helix-turn-helix motif is located in the third quarter of most of the
sequences; the N-terminal and central regions of these proteins are presumed
to interact with effector molecules and may be involved in dimerization [3].
The minimal DNA binding domain, which spans roughly 100 residues and comprises
```

Figure 1. Sample data from PROSITE.

High specificity. In the majority of cases we have chosen patterns or profiles that are specific enough that they do not detect too many unrelated sequences, yet they will detect most, if not all, sequences that clearly belong to the set in consideration.

Documentation. Each of the entries in PROSITE is fully documented; the documentation includes a concise description of the protein family or domain that it is designed to detect as well as a summary of the reasons leading to the development of the pattern or profile.

Periodic reviewing. It is important that each entry be periodically reviewed to ensure that it is still valid.

A very tight relationship with the SWISS-PROT protein sequence data bank (12). Updating of PROSITE and of the annotations of the relevant SWISS-PROT entries are very often done in parallel. Software tools based on PROSITE are used to automatically update the feature table lines of SWISS-PROT entries relevant to the presence and extent of specific domains.

FORMAT AND DOCUMENT FILES

The core of the PROSITE database is composed of two ASCII (text) files. The first file (PROSITE.DAT) is a computer-readable file that contains all the information necessary for programs that make use of PROSITE to scan sequence(s) for the occurrence of the patterns and/or profiles. This file also includes, for each entry described, statistics on the number of hits obtained while scanning for that pattern or profile in SWISS-PROT. Cross-references to the corresponding SWISS-PROT entries are also present in the file. The second file (PROSITE.DOC), which we call the

```
the HTH motif contains another region with similarity to classical HTH domain.
However, it contains an insertion of one residue in the turn-region.

A signature pattern was derived from the region that follows the first HTH
domain and that includes the totality of the putative second HTH domain. A
more sensitive detection of members of the araC family is available through
the use of a profile which spans the minimal DNA-binding region of 100
residues

 Consensus pattern: [KRQ]-[LIVMA]-x(2)-[GSTALIV]-[FYWGDH]-x(2)-[LIVMSA]-
                    x(4,5)-[LIVMF]-x(2)-[LIVMSTA]-[GSTACTL]-x(3)-[GAPURF]-
                    [LIVMFT]-x(4,5)-[LFT]-x(3)-[FYIVA]-[FYGHCW]-x(3)
                   [GSADENQKR]-x-[DNSTAPKL]-[PARL]
-Sequences known to belong to this class detected by the pattern: ALL.
-Other sequence(s) detected in SWISS-PROT: 34.

 Sequences known to belong to this class detected by the profile: ALL.
-Other sequence(s) detected in SWISS-PROT: NONE.

-Note: this documentation entry is linked to both a signature pattern and a
 profile. As the profile is much more sensitive than the pattern, you should
 use it if you have access to the necessary software tools to do so.

 Expert(s) to contact by email:
       Ramos J.L.; jlramos@samba.cnb.uam.es
       Gallegos M. T.; mtrivi@samba.cnb.uam.es

-Last update: November 1997 / Text revised.

[ 1] Gallegos M.-T., Michan C., Ramos J.L.
     Nucleic Acids Res. 21:807-810(1993).
[ 2] Henikoff S., Wallace J.C., Brown J.P.
     Meth. Enzymol. 183:111-132(1990).
[ 3] Bairoch A., Schleif R.F.
     Proc. Natl. Acad. Sci. U.S.A. 90:5636-5642(1993).

| This PROSITE entry is copyright by the Swiss Institute of Bioinformatics |
| (SIB). There are no restrictions on its use by non-profit institutions as |
| long as its content is in no way modified and this statement is not       |
| removed. Usage by and for commercial entities requires a license agreement |
| (See http://www.isb-sib.ch/announce/ or email to license@isb-sib.ch).     |

(END)
```

1b) The corresponding pattern and profile entries in the PROSITE.DAT file

```
ID   HTH_ARAC_FAMILY_1; PATTERN
AC   PS01041;
DT   APR 1990 (CREATED); NOV-1995 (DATA UPDATE); JUL 1998 (INFO UPDATE).
DE   Bacterial regulatory proteins, araC family signature.
PA   [KRQ]-[LIVMA]-x(2)-[GSTALIV]-[FYWGDH]-x(2)-[LIVMSA]-x(4,5)-[LIVMF]-
PA   x(2)-[LIVMSTA]-[GSTACTL]-x(3)-[GAPURF]-[LIVMFT]-x(4,5)-[LFT]-x(3)-
PA   [FYIVA]-[FYWHCH]-x(3)-[GSADENQKR]-x-[DNSTAPKL]-[PARL].
NR   /RELEASE=36,74019;
NR   /TOTAL=1154(107); /POSITIVE=78(73); /UNKNOWN=0(0); /FALSE_POS=37(34);
NR   /FALSE_NEG=8; /PARTIAL=0;
CC   /TAXO-RANGE=??EP?; /MAX_REPEAT=1;
DR   P43463, AARP_PROST, T; P19219, ADAA_BACSU, T; Q10630, ADA_MYCTU , T;
DR   P33234, ADIY_ECOLI, T; P43466, AGUR_ECOLI, T; P05052, APPY_ECOLI, T;
DR   P11765, ARAC_CITFR, T; P03021, ARAC_ECOLI, T; P07642, ARAC_ERWCH, T;
DR   P03022, ARAC_SALTY, T; Q05320, ARAL_STRAT, T; P35319, ARAL_STRLI, T;
DR   P17410, CELD_ECOLI, T; P25393, CFAD_ECOLI, T; P43460, CSVR_ECOLI, T;
DR   P10854, ENVY_ECOLI, T; P26993, EXSA_PSEAE, T; P23774, FAPR_ECOLI, T;
DR   P31778, HRPB_BURSO, T; P19437, INVF_SALTY, T; P28808, LCRF_YERPE, T;
DR   Q51872, LUXO_PHOLE, T; P27246, MARA_ECOLI, T; Q56070, MARA_SALTY, T;
DR   P10411, MELR_ECOLI, T; P28809, MMSR_PSEAE, T; Q00753, MSHR_ATRHU, T;
DR   Q04642, MXIE_SHIFL, T; Q55291, MXIE_SHISO, T; P40883, PCHR_PSEAE, T;
DR   P43459, PDRA_ECOLI, T; Q05587, POCR_SALTY, T; Q52620, PQRA_PROVU, T;
DR   P43465, RAFR_PHYPE, T; Q08813, RAMA_KLEPN, T; P55178, RHAR_ECOLI, T;
DR   P40865, RHAR_SALTY, T; P09377, RHAS_ECOLI, T; P27029, RHAS_SALTY, T;
DR   P16114, RNS_ECOLI , T; P22539, SOXS_ECOLI, T; Q56143, SOXS_SALTY, T;
DR   P29452, TCPN_VIBCH, T; P43462, THCR_RHOSO, T; P32326, URER_ECOLI, T;
DR   Q02458, VIRF_PRONI, T; Q04248, VIRF_SHIDY, T; P13225, VIRF_YEREN, T;
DR   P37390, XYLR_ECOLI, T; P45043, XYLR_HAEIN, T; P07859, XYLS_PSEPU, T;
DR   Q04710, XYS1_PSEPU, T; Q05092, XYS2_PSEPU, T; Q05335, XYS3_PSEPU, T;
DR   Q04713, XYS4_PSEPU, T; P55449, Y4PX_RHIEN, T; P45008, YA52_HAEIN, T;
DR   P40409, YBBR_BACSU, T; P43462, YCGR_ALTCA, T; P76241, YEAM_ECOLI, T;
DR   P36547, YFEO_ECOLI, T; P54722, YFIF_BACSU, T; P37638, YHIW_ECOLI, T;
DR   P37639, YHIX_ECOLI, T; P31449, YIDL_ECOLI, T; P32677, YIJO_ECOLI, T;
DR   P40331, YLXR_BACSU, T; P45418, YMCR_STRLA, T; P06134, ADA_ECOLI , T;
DR   P26189, ADA_SALTY , T; P26950, CAFR_YERPE, T; P27292, ROB_ECOLI , T;
DR   P28816, TETD_ECOLI, T;
DR   Q47129, FEAR_ECOLI, N; P55922, AAMA_RHYCL, N; Q06861, V38K_MYCTU, N;
DR   P77634, YBCM_ECOLI, N; P77601, YKOA_ECOLI, N; P77379, YKGD_ECOLI, N;
DR   Q46855, YQHC_ECOLI, N; P71663, YK17_MYCTU, N;
DR   P28647, AAJR_RAT  , F; P74985, ARSD_YEREN, F; Q54468, CKB_SERMA , F;
DR   P23577, CYP_CHLRE , F; P43712, FASD_HAEIN, F; Q44763, PLIP_BORBU, F;
DR   P74930, PLIP_TREPA, F; O42101, FTF_CHICK , F; Q04952, GLS3_YEAST, F;
DR   P23970, HEMD_BACSU, F; O15303, HUR6_HUMAN, F; P35369, HUR6_RAT  , F;
DR   P40931, HPL_EPLV  , F; P55015, HXC2_RABIT, F; P55016, HXC2_RAT  , F;
DR   P29801, HU2C_STYPI, F; P72714, HU2C_SYNY3, F; P28531, KLS_CHLTR , F;
DR   P31983, RP54_ACICA, F; P54991, SNAA_STRPR, F; P40238, TPOR_HUMAN, F;
DR   Q58351, TPOR_MONBE, F; P23626, V3A_TAV   , F; P15911, VFP3_POMPV, F;
DR   P47571, Y129_MYCGE, F; P77400, YBAT_ECOLI, F; P75826, YBJE_ECOLI, F;
DR   P77744, YDAX_ECOLI, F; P87293, YDN1_SCHPO, F; Q22288, Y179_ARATH, F;
DR   Q63106, YY07_METJA, F; Q45980, FLIP_CAUCR, F; P55719, Y4YK_RHISN, F;
DR   P29960, YCB7_PSEPS, F;
DO   PDOC00040;
//
```

Figure 1. *continued*

textbook, contains textual information that documents each pattern.

A sample textbook entry is shown (Fig. 1a); this particular entry is linked to two entries in the PROSITE.DAT file: a pattern and a profile (Fig. 1b).

Several document files are also distributed with the database:

PROSUSER.TXT	The database user's manual
PROFILE.TXT	A detailed description of the syntax for the profiles
PROSITE.LIS	A list of PROSITE documentation entries
PROSITE.GET	A document on how to obtain a local copy of PROSITE
PROSITE.PRG	A description of programs and electronic mail servers that make use of PROSITE
PAUTINDX.TXT	An index of authors cited in the PROSITE.DOC file

CONTENT OF THE CURRENT RELEASE

Release 15.0 of PROSITE (July 1998) contains 1014 documentation entries describing 1352 different patterns, rules and profiles/matrices. In addition to these entries, a collection of 241 preliminary profiles is available in the pre-release distribution from the FTP server of the ISREC group (see below). The list of the documentation entries that have been added since the last release of PROSITE (14.0) is provided in Table 1, furthermore, many entries were updated. The database requires ~5 Mb of disk storage space. The present distribution frequency is two releases per year. No restrictions are placed on use or redistribution of the data. Future releases of PROSITE will be copyright (releases up to number 15.0 are not).

HOW TO OBTAIN A LOCAL COPY OF PROSITE

By CD-ROM

PROSITE is distributed on CD-ROM by the EMBL Outstation—the European Bioinformatics Institute (EBI) (13). For all enquiries regarding the subscription and distribution of PROSITE one should contact: The EMBL Outstation—The European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Tel: +44 1223 494 444; Fax: +44 1223 494 468; Email: datalib@ebi.ac.uk

```
ID   HTH_ARAC_FAMILY_2; MATRIX
AC   PS01124;
DT   NOV 1995 (CREATED); NOV 1995 (DATA UPDATE); JUL 1998 (INFO UPDATE)
DE   Bacterial regulatory proteins, araC family DNA-binding domain profile.
MA   /GENERAL_SPEC: ALPHABET='ABCDEFGHIKLMNPQRSTVWYZ'; LENGTH=99;
MA   /DISJOINT: DEFINITION=PROTECT; NI=6; NJ=94;
MA   /NORMALIZATION: MODE=1; FUNCTION=LINEAR; R1=1.5162; R2=0.0218; TEXT='OrigScore';
MA   /CUT_OFF: LEVEL=0; SCORE=120; N_SCORE=8.5; MODE=1;
MA   /DEFAULT: D=-20; I=-20; B1= 70; I1=-70; M1=-105; M0=-105; IM=-105; DM= 105;
MA   /I: B1=0; B1=-105; D0=-105;
MA   /M: SY='D'; M=-10.11,-25,14.11,-25,-12,2, 25,4, 22, 15,7,-13,8,4,0,-7,-23,-25,-15,10;
MA   /M: SY='R'; M=-7,-1,-26,-1,5, 24, 15,-1,-20,15, 19,-11,1,-11,8,20,-3, 6, 18, 22,-11,5;
MA   /M: SY='V'; M=8,-24,-17,-29, 22, 2, 24,-25,21, 21,14,9, 22,-22,-20,-22,-11,-3,22,-23,
8,-22;
MA   /M: SY='V'; M=-7,-18,-10,-21, 13, 7,-25,-14,6, 11,5,4,-15,-23,-10, 5, 11, 2,6,-23,-5,-
13;
MA   /M: SY='Q'; M=-2,-1, 20,-3,2,-23,-10,1, 19,0, 14,-7,0,-15,7,3,-1,-4,-16,-26, 12,3;
..
...
.... Lot of lines omitted.
...
..
MA   /M: SY='R'; M= 13, 14,-25,-15,-6,-9,-21,-7,-13,8, 3, 2, 7, 21,0,32,-11,-4, 9,-20,-6,-6;
MA   /M: SY='R'; M= 4, 4, 26, 6,2,-20,-16,-1,-17,9,-14, 6, 1, 15,7,11, 3,-4, 14, 22, 9,3;
MA   /M: SY='P'; M=-3,-7,-26,-6,-5,-17,-10, 2, 13,0, 12,-4,-2,-17,3,5, 4,-5,-12,-21, 6, 2;
MA   /I: B1=0; I1= 105; D2=-105;
MA   /RELEASE=36,78019;
MA   /TOTAL=81(81); /POSITIVE=81(81); /UNKNOWN=0(0); /FALSE_POS=0(0);
MA   /FALSE_NEG=0; /PARTIAL=0;
CC   /TAXO-RANGE=????P?; /MAX-REPEAT=1;
DR   P03463, AARP_PROST, T; P19219, ADHA_BACSU, T; P06134, ADA_ECOLI , T;
DR   Q10630, ADA_MYCTU , T; P26189, ADA_RALTY , T; P33234, AUIY_ECOLI, T;
DR   P43464, AGRP_ECOLI, T; P05052, APPY_ECOLI, T; P11769, ARAC_CITFR, T;
..
...
.... Lot of lines omitted.
...
..
DR   P37639, YHJX_ECOLI, T; P31449, YIDL_ECOLI, T; P32677, YIJD_ECOLI, T;
DR   P40331, YISR_BACSU, T; P77601, YKGA_ECOLI, T; P77375, YBCD_ECOLI, T;
DR   P41458, YHCR_STRLA, T; Q46855, YOHC_ECOLI, T; P71663, YR13_MYCTU, T;
3D   1ADH; 1SFE; 2AAC; 2AHA; 2ARC;
DO   PDOC00040;
//
```

Figure 1. *continued*

Table 1. List of patterns documentation entries that have been added since the last release of PROSITE (14.0)

DNA repair protein radC family signature
recR protein signature
ubiH/COQ6 monooxygenase family signature
ATP phosphoribosyltransferase signature
Prolipoprotein diacylglyceryl transferase signature
Phosphatidate cytidylyltransferase signature
Lipoate-protein ligase B signature
moaA / nifB / pqqE family signature
BCCT family of transporters signature
Flagellar motor protein motA family signature
Protein secA signatures
ATP1G1 / PLM / MAT8 family signature
Protein smpB signature
Uncharacterized protein family UPF0044 signature
Uncharacterized protein family UPF0047 signature
Uncharacterized protein family UPF0054 signature
Uncharacterized protein family UPF0057 signature

By anonymous FTP

If you have access to a computer system linked to the Internet you can obtain PROSITE using FTP (File Transfer Protocol), from the following file servers:

ExPASy (Expert Protein Analysis System) server, Swiss Institute of Bioinformatics (SIB); Internet address: ftp://www.expasy.ch/databases/prosite/

ISREC (Swiss Institute for Experimental Cancer Research) anonymous FTP server, Swiss Institute of Bioinformatics (SIB); Internet address: ftp://ftp.isrec.isb-sib.ch/sib-isrec/profiles/

EBI (European Bioinformatics Institute) anonymous FTP server; Internet address: ftp://ftp.ebi.ac.uk/pub/databases/prosite/

The pre-release collection of profiles is only available from the ISREC FTP server.

By Email through the EBI network fileserver

PROSITE can be obtained from the EBI network fileserver. Detailed instructions on how to make the best use of this service, and in particular on how to obtain PROSITE, can be obtained by sending to the network address netserv@ebi.ac.uk the following message:

HELP
HELP PROSITE

HOW TO MAKE USE OF PROSITE

Computer programs

Many academic groups and commercial companies have developed computer programs that make use of the pattern entries in PROSITE. The 'PROSITE.PRG' file contains a full list of these programs, their operating system specificity, characteristics as well as information on how to obtain them.

Two software packages are distributed to make use of profile entries:

(i) *pftools* (version 2.1 in FORTRAN77) written by Philipp Bucher. pfscan loads a sequence from a file and scans it with all (or one) of PROSITE profiles; pfsearch loads a profile from a file and scans for it in a SWISS-PROT database file. These tools are available by anonymous FTP from the server: ftp://ftp.isrec.isb-sib.ch/sib-isrec/pftools . Several versions are available, as well as executables compiled for many unix platforms and for Windows 95/98.

(ii) *PrfLib* (version 1.0 in ANSI C) written by Nicolas Moeri. scan4prf loads a sequence from a file and scans it with all (or one) of PROSITE profiles; srch4prf loads a profile from a file and scans for it in a SWISS-PROT database file. These tools are available from the server: http://mamac29.epfl.ch/

Email servers

There are many Email servers that are available to molecular biologists (14). This an example of a server taking advantage of the PROSITE database:

Name:	MOTIF E-Mail Server on GenomeNet
Organization:	Supercomputer Laboratory, Kyoto Institute for Chemical Research, Japan
Description:	Allows to rapidly compare a new protein sequence against all patterns stored in PROSITE as well as in the MotifDic library (15).
Server email address:	motif@genome.ad.jp
Address to report problems:	motif-manager@genome.ad.jp

Interactive access to PROSITE using the World Wide Web

The most efficient and user-friendly way to browse interactively in PROSITE as well as to analyze a sequence for the occurrence of a pattern or a profile is to use the World-Wide Web (WWW) molecular biology server ExPASy (16). Using a WWW browser, one has access to all the hypertext documents stored on the ExPASy server (as well as many other WWW servers) and also can make use of many sequence analysis software tools.

The ExPASy server may be accessed through its URL which is: http://www.expasy.ch/ . You can directly access to the 'top' page of the section of ExPASy that allows you to browse through the PROSITE documentation and data entries by opening the URL: http://www.expasy.ch/sprot/prosite.html

To use the PROSITE patterns and profiles, you can make use of the following software tools.

ScanProsite. Allows the user to either scan a protein sequence—from SWISS-PROT or provided by the user—for the occurrence of patterns stored in PROSITE or to scan the SWISS-PROT and/or TrEMBL database—including weekly releases—for the occurrence of a pattern that can originate from PROSITE or be provided by the user. The URL for ScanProsite is: http://www.expasy.ch/sprot/scnpsite.html

ProfileScan. Allows the user to scan a protein sequence—from SWISS-PROT or provided by the user—for the occurrence of profiles stored in PROSITE. The URL for ProfileScan is: http://www.isrec.isb-sib.ch/software/PFSCAN_form.html

FrameProfileScan. Allows the user to scan a DNA sequence (translated on the fly into protein)—from EMBL or provided by the user—for the occurrence of profiles stored in PROSITE. The URL for FrameProfileScan is: http://www.isrec.isb-sib.ch/software/PFRAMESCAN_form.html

REFERENCES

1 Bairoch,A. and Bucher,P. (1994) *Nucleic Acids Res.*, 22, 3583–3589.
2 Bairoch,A., Bucher,P. and Hofmann,K. (1997) *Nucleic Acids Res.*, 25, 217–221.
3 Doolittle,R.F. (1986) *Of URFs and ORFs: A Primer On How To Analyze Derived Amino Acid Sequences*. University Science Books, Mill Valley, California.
4 Lesk,A.M. (1988) In Lesk,A.M. (ed.), *Computational Molecular Biology*. Oxford University Press, Oxford, pp. 17–26.
5 Bucher,P. and Bairoch,A. (1994) In Altman,R., Brutlag,D., Karp,P., Lathrop,R. and Searls,D. (eds), *ISMB-94; Proceedings Second International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, pp. 53–61.
6 Bucher,P., Karplus,K., Moeri,N. and Hofmann,K. (1996) *Comput. Chem.*, 20, 3–23.
7 Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl Acad. Sci. USA*, 84, 4355–4358.
8 Eddy,S.R. (1996) *Curr. Opin. Struct. Biol.*, 6, 361–365.
9 Gribskov,M., Luethy,R. and Eisenberg,D. (1990) *Methods Enzymol.*, 183, 146–159.
10 Luethy,R., Xenarios,I. and Bucher,P. (1994) *Protein Sci.*, 3, 139–146.
11 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Comput. Applic. Biosci.*, 10, 19–29.
12 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res.*, 26, 38–42.
13 Stoesser,G., Moseley,M.A., Sleep,J., McGowran,M., Garcia-Pastor,M. and Sterk,P. (1998) *Nucleic Acids Res.*, 26, 8–15.
14 Henikoff,S. (1993) *Trends Biochem. Sci.*, 18, 267–268.
15 Ogiwara,A., Uchiyama,I., Seto,Y. and Kanehisa,M. (1992) *Protein Engng*, 5, 479–488.
16 Appel,R.D., Bairoch,A. and Hochstrasser,D.F. (1994) *Trends Biochem. Sci.*, 19, 258–260.

Journal of Molecular Biology
Volume 225, Issue 2 , 20 May 1992, Pages 487-494

doi:10.1016/0022-2836(92)90934-C Cite or Link Using DOI

Article

Membrane protein structure prediction *1

Hydrophobicity analysis and the positive-inside rule

Gunnar von Heijne

Department of Molecular Biology Karolinska Institute Center for Structural Biochemistry NOVUM, S-141 57, Huddinge, Sweden

Received 14 October 1991; accepted 20 January 1992. Available online 28 October 2004.

Abstract

A new strategy for predicting the topology of bacterial inner membrane proteins is proposed on the basis of hydrophobicity analysis, automatic generation of a set of possible topologies and ranking of these according to the positive-inside rule. A straightforward implementation with no attempts at optimization predicts the correct topology for 23 out of 24 inner membrane proteins with experimentally determined topologies, and correctly identifies 135 transmembrane segments with only one overprediction.

Author Keywords: membrane protein; protein structure; prediction

*1 This work was supported by grants from the Swedish Natural Sciences Research Council and the Swedish Board for Technical Development.

Journal of Molecular Biology
Volume 225, Issue 2 , 20 May 1992, Pages 487-494

This Document
- **Abstract**
- Abstract + References
- PDF (1194 K)

Actions

1: Eur J Biochem. 1993 May 1;213(3):1333-40.

Predicting the topology of eukaryotic membrane proteins.

Sipos L, von Heijne G.

Department of Theoretical Physics, Royal Institute of Technology, Stockholm, Sweden.

We show that the so-called 'positive inside' rule, i.e. the observation that positively charged amino acids tend to be more prevalent in cytoplasmic than in extra-cytoplasmic segments in transmembrane proteins [von Heijne, G. (1986) EMBO J. 5, 3021-3027], seems to hold for all polar segments in multi-spanning eukaryotic membrane proteins irrespective of their position in the sequence and hence can be used in conjunction with hydrophobicity analysis to predict their transmembrane topology. Further, as suggested by others, we confirm that the net charge difference across the first transmembrane segment correlates well with its orientation [Hartmann, E., Rapoport, T. A. and Lodish H. F. (1989) Proc. Natl Acad. Sci. USA 86, 5786-5790], and that the overall amino-acid composition of long polar segments can also be used to predict their cytoplasmic or extra-cytoplasmic location [Nakashima, H. and Nishikawa, K. (1992) FEBS Lett. 303, 141-146]. We present an approach to the topology prediction problem for eukaryotic membrane proteins based on a combination of these methods.

MeSH Terms:

- Aspartic Acid/analysis
- Glutamates/analysis
- Glutamic Acid
- Membrane Proteins/analysis
- Membrane Proteins/chemistry*
- Research Support, Non-U.S. Gov't
- Tryptophan/analysis
- Tyrosine/analysis

Substances:

- Glutamates
- Membrane Proteins
- Tyrosine
- Aspartic Acid
- Glutamic Acid

TMpred - Prediction of Transmembrane Regions and Orientation

The TMpred program makes a prediction of membrane-spanning regions and their orientation. The algorithm is based on the statistical analysis of TMbase, a database of naturally occuring transmembrane proteins. The prediction is made using a combination of several weight-matrices for scoring.

K. Hofmann & W. Stoffel (1993)
TMbase - A database of membrane spanning proteins segments
Biol. Chem. Hoppe-Seyler **374**,166

For further information see the TMbase and TMpredict documentation.

Usage: Paste your sequence in one of the supported formats into the sequence field below
and press the "Run TMpred" button.
Make sure that the format button (next to the sequence field) shows the correct format

Choose the minimal and maximal length of the hydrophic part of the transmembrane helix

Output format: html | minimum: 17 | maximum: 33

Query title (optional):

Input sequence format: Plain Text

Query sequence: or ID or AC or GI (see above for valid formats):

Run TMpred | Clear Input

Go back to the EMBnet.ch home page

MF C-35 A Database of Membrane Spanning Protein Segments

K. Hofmann and W. Stoffel
Institut fur Biochemie, Medizinische Fakultät, Universitat zu Köln, Köln, FRG

A database of all protein segments that are reported to span a membrane has been extracted from SwissProt 22. This sub-database consists of several tables that can be used with any relational database management system. The information stored within the database contains besides the sequence itself both annotational items extracted from SwissProt and additional data fields calculated from the sequence or taken from other sources. Important data fields include, for example, the putative transmembrane sequence, the sequence of the flanking regions, taxonomic information, the presumed orientation of the segment, calculated values for hydrophobicity and hydrophobic moment, and grouping into families by either functional or sequence relatedness of the proteins.
This database together with a set of related programs has been used to analyze the presumed transmembrane segments for positional preferences of amino acid residues. The influences of neighbouring residues, membrane protein classification, taxonomic classification and segment orientation on these positional preferences have been studied.

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☑ FADED TEXT OR DRAWING
- ☑ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.